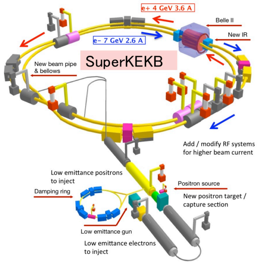
HUF 2016 KEK SITE REPORT

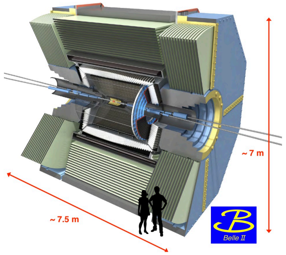
Report on HPSS/GHI Site Migration to New KEKCC

Koichi Murakami (KEK/CRC)

HUF 2016 NYC

KEK ON-GOING PROJECTS


e+ 4 GeV 3.6 A
Belle II
New IR
e− 7 GeV 2.6 A
SuperKEKB
New beam pipe & bellows
Add / modify RF systems for higher beam current
Low emittance positrons to inject
Damping ring
Positron source
New positron target / capture section
Low emittance gun
Low emittance electrons to inject



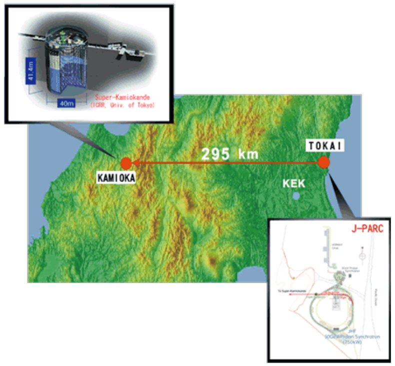
BELLE, BELLE II EXPERIMENTS

Belle experiment, precise measurements for CP violation.

Belle II is the next generation Belle experiment. Aim to discover new physics beyond the SM.

Physics run will start in 2017.

T2K



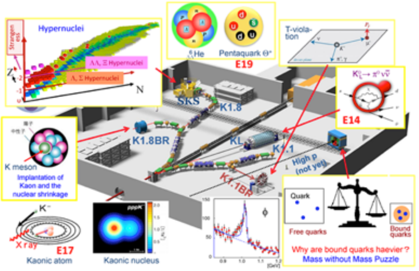
Neutrino experiment for measuring neutrino mass and flavour mixing.

Shoot neutrino from Tokai to the detector at Kamioka mine (300km away)

HADRON EXPERIMENTS AT J-PARC

Various experiments for kaon and hadron physics

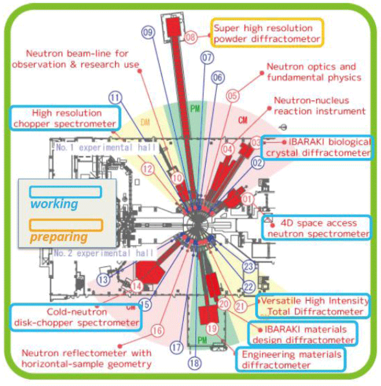
Nuclear & Particle Physics Program at J-PARC Hadron Experimental Hall



MATERIAL AND LIFE SCIENCE AT J-PARC

Neutron diffraction, neutron spectroscopy,

nano-structure analysis, neutron instruments, muon spectroscopy


Super high resolution powder diffractometor
Neutron beam-line for observation & research use
Neutron optics and fundamental physics
Neutron-nucleus reaction instrument
High resolution chopper spectrometer
IBARAKI biological crystal diffractometer
No. 1 experimental hall
working
preparing
No. 2 experimental hall
4D space access neutron spectrometer
Versatile High Intensity Total Diffractometer
Cold-neutron disk-chopper spectrometer
IBARAKI materials design diffractometer
Neutron reflectometer with horizontal-sample geometry
Engineering materials diffractometer
DM
PM
CM
01
02
03
04
05
06
07
08
09
10
11
12
13
14
15
16
17
18
19
20
21
22
23

KEKCC System Replacement

System is totally replaced every 4-5 years, according to Japanese government procurement rule for computer system.

- Not in-house scale-out model, but rental system
- Purchase including system operation
- Completely different purchase/operation model from US/EU sites
- Last system replacement in Feb/2012

System implementation (Jan – Aug / 2016)

- Facility updates (power supply, cooling)
- Hardware installation
- System design / implementation / testing

The new system will be released at Sep/01/2016.

lenovo
Mellanox TECHNOLOGIES
lenovo
DDN
DataDirect NETWORKS
IBM
IBM

NEW KEKCC

CURRENT VS NEXT

	Current	New	Upgrade Factor
CPU Server	IBM iDataPlex	Lenovo NextScale	
CPU	Xeon 5670 (2.93 GHz ,6core)	Xeon E5-2697v3 (2.6GHz, 14cores)	
CPU cores	4,080	10,024	**x2.5**
IB	QLogic 4xQDR	Mellanox 4xFDR	
Disk Storage	DDN SFA10K 4 PB	IBM Elastic Storage System (ESS) 10PB	
HSM Disk Storage	DDN SFA10K 3PB	DDN SFA12K 3PB	
Disk Capacity	7 PB (3PB for HSM)	13 PB (3PB for HSM)	**x1.8**
Tape Library	TS3500 (12 racks)	TS3500 (13 racks)	
Tape Drive	IBM TS1140 x 60	IBM TS1150 x54	
Tape max capacity	16 PB	70 PB	**x4.3**

HSM SYSTEM

HPSS/GHI servers

IBM

TS3500

DataDirect NETWORKS

DDN SFA 12K

GPFS (GHI) : 3PB
Total throughput : > 50 GB/s

TAPE SYSTEM

TAPE LIBRARY

- IBM TS3500 (13 racks)
- Max. capacity : 70 PB

TAPE DRIVE

- TS1150 : 54 drives
- TS1140 : 12 drives (for media conversion)
- We do not use LTO.

IBM TS3500

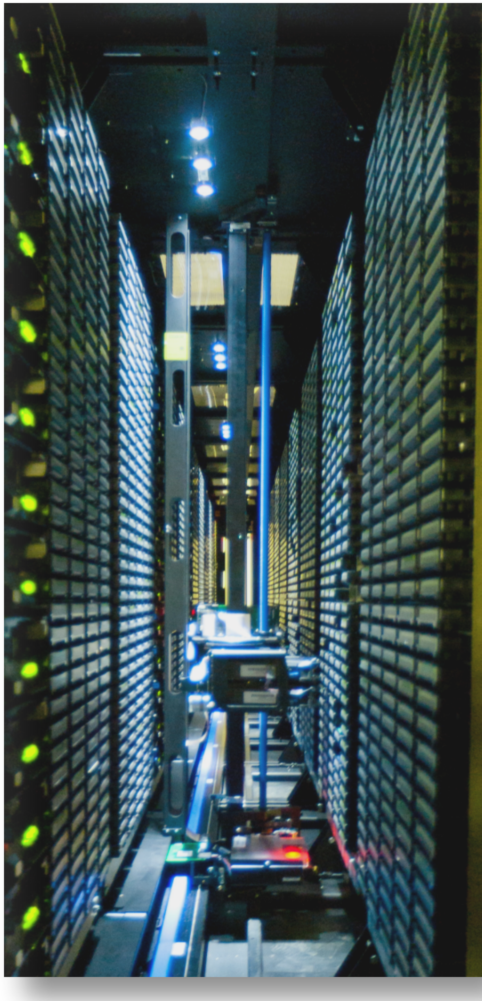
TAPE MEDIA

- JD : 10TB, 360 MB/s
- JC : 7TB, 300 MB/s (reformatted)
  - Reformatting will be done in background for 6-12 months (expected).
- JC : 4TB, 250 MB/s
- Users (experiment groups) pay tape media they use.

GHI, GPFS + HPSS :
THE BEST OF BOTH WORLDS

HPSS

High Performance Storage System



HPSS

- □ We have used HPSS as HSM system for last 15+ years.
- □ 1st layer : GGPS DDN 3PB + 2nd layer : IBM Tape

GHI, GPFS + HPSS

- □ GPFS parallel file system staging area
- □ Perfect coherence with GPFS access (POSIX I/O)
- □ KEKCC is the pioneer of GHI customers (since 2012).
- □ Data access with high I/O performance and good usability.
  - □ Same access speed as GPFS, once data staged
  - □ No HPSS client API, no changes in user codes
  - □ small file aggregation helps tape performance for small data

NEW SYSTEM CONFIGURATION PARAMS.

Software	Current	New
HPSS	7.3.3.9	7.4.3.2
GHI	2.3.1.2	2.5.0.1
GPFS	3.5.0.18	4.2.0.1
OS (HPSS nodes)	RHEL 5.9	RHEL 6.7
OS (GHI nodes)	RHEL 5.6	RHEL7.1

Component	Qty.
HPSS Core Server	**1**
HPSS Disk Mover	4
HPSS Tape Mover	3
Mover Storage	600 TB
Max. #Files	**2 Billion** (x10)
GHI IOM	6
GHI Session Server	3

HPSS core server is not a single point of failure for us.

SYSTEM MIGRATION WORKS

HSM service on the current system

- □ 3-days downtime for system migration (backup of the current / restore in the new)
- □ Keep GPFS disk mount (read-only) for 2 weeks before the new system
  - □ Only staged data on disk is accessible.

System migration

- □ 8.5 PB data, 170 M files, 5,000 tapes
- □ 3-days work on Aug / 15 – 17
  - □ Move physical tapes from the current to new tape library
  - □ DB2 migration using QRep
  - □ GHI backup and restore

Take checksum for tape data

- □ 6 months work for higher priority data
- □ Taken directly from tapes (tape-ordered, htared file for small files, as hpss file)
  - □ 200 MB/s in average, 4,000 vols.
- □ Store checksum and timestamp into GPFS UDA

SYSTEM IMPROVEMENTS (1)

Separate GPFS clusters

- □ GPFS disk system (10PB) and GHI GPFS system (3PB)
- □ ITO stability and system management (maintenance, updates,..)

Introduce GPFS local cache as layer-0 disk

- □ for SSD of batch servers
  - □ <4GB files cached in local SSD
- □ reducing concurrent access to GPFS files from many clients

COS supports mixed media types

- □ Can mix different types of tape media as **RW COS**
- □ JB/JC/JD

Do not purge small files (<8MB)

- □ number of small files is too big, but no impact on disk space

SYSTEM IMPROVEMENTS (2)

Improve migration way

- □ current : listing all migration files, then migrate one time:
  - □ Single migration requests for >100 k files overflows the hpss queues, migration stalled.
- □ new : migration by 10k files in ghi_backup

Bulk staging & Coherent mechanism with batch job scheduler

- □ Tons of staging requests results in much waiting time.
- □ Tape-ordered bulk staging is efficient. Automated way from a recall list.
- □ Data staging before dispatching jobs. Coherent efforts with LSF

GHI issues

- □ CR-35: Conversion from HPSS file name to GHI file name (GHI 2.5)
  - □ Search orphan files in GHI (created by the old bug in htar)
- □ An alternative CR-12: Repack HTAR files (GHI 2.5)

HPSS / GHI PERFORMANCE MEASUREMENTS

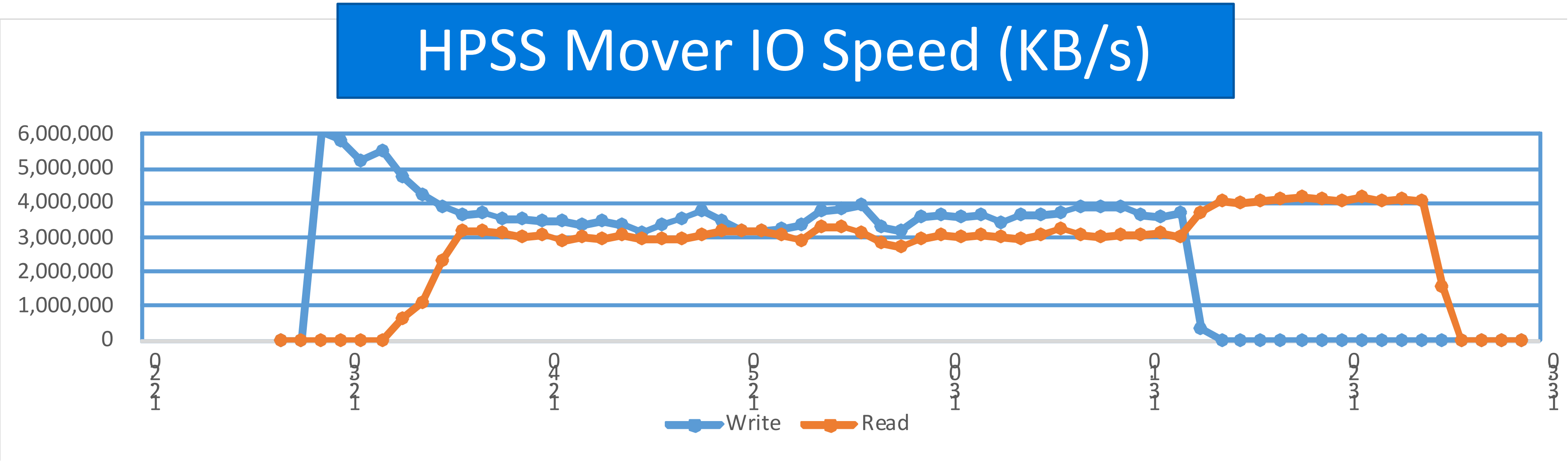
REQUIREMENTS :

- □ Max. expected data writing (sustained) / migration : 200 TB / day
- □ Max. expected staging : 50 TB / day
- □ Requirements from experiments

MEASUREMENTS :

- □ Mover IO : 3 GB/s (read / write)
- □ Migration (via ghi_backup) :
  - □ 3.4 GB/s (4GB, 24p), > 200 TB / day
- □ Staging :
  - □ > 100 TB / day (1GB, tape-order, >1.2GB/s, 8p)
  - □ 20 TB / day (2GB, non-tape-order, 0.25 GB/s, 8p)
- □ Staging & Migration :
  - □ 0.2 GB/s staging & 2.4 GB/s migration (2GB, non-tape-order, 24p)

HPSS MOVER IO PERFORMANCE



MOVER IO PERFORMANCE:

- □ Data : migration (ghi_backup) for 2 GB files
- □ Write : GHI migration / Read : HPSS migration
- □ aggregate for nmmon outputs of 4 movers
- □ 3 GB/s for concurrent read / write accesses

FUTURE : DATA EXPLOSION

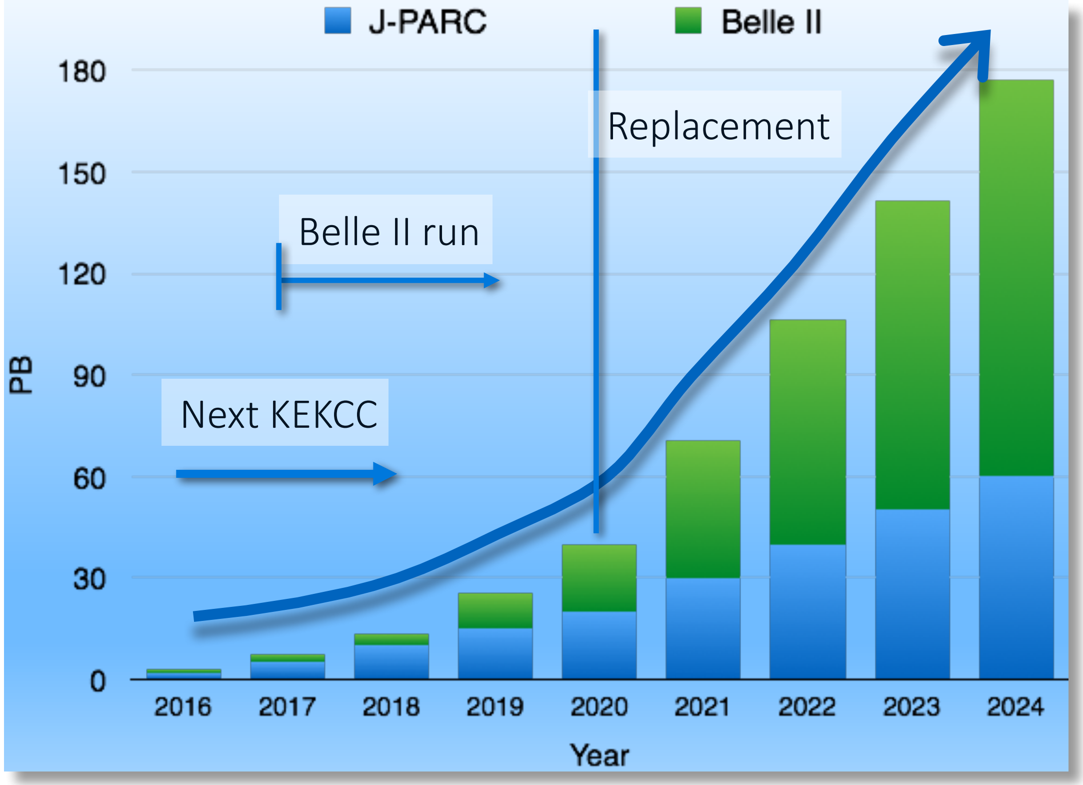
THIS YEAR

- □ 8.5 PB, 170 M files
  +2.5PB / year

DATA GROWTH EXPECTATION

- □ J-PARC will constantly produce data. A few – 10 PB /year
- □ Data explosion is expected for Belle II.
- □ Data growth rate beyond 2020 is very high.

Data Explosion



CHALLENGES TO THE FUTURE

Concerns on migration

- Our System will be replaced every 4-5 years.
- Expected amount of data migration
  - 8PB (2016), Nx10PB (2020), Nx100PB (2024)
- Migration issues will be critical.

Challenges on scalability of the system

- How to scale out the system to Exascale
- Coherent data management
  - Data taking, archive, processing, preservation…
- Monitoring & Visualize system healthy :
  - We are monitoring some resources usage (I/O, tape drive usage,...)
  - Experience difficulty of identify problems
  - Elasticseach and Kibana can help?

SUMMARY

- Next KEKCC system will start in September 2016.
  - Increase computing resources :
    CPU : 10K cores (x2.5), Disk : 13PB (x1.8), Tape : 70PB (x4.3)
- HPSS/GHI system migration was well done.
  - Minum service down-time
  - Performance as designed values
  - Impronents on system operation
  - Thanks for our collaobrative work with IBM team
- Scalable data management is a challenge for next 10 years.
  - Data explosion is expected as Belle II experiment will start in 2017.
  - Data processing cycle (data taking, archive, processing, preservation…)